

Språkbanken och Korp: Mot en språkteknologibaserad forskningsinfrastruktur

Lars Borin
Språkbanken/svenska språket, Göteborgs universitet
Giellatekno, UiT

19/2 2014

- ~1970: första svenska korpusen: Press-65
- 1972: professur i språkvetenskaplig databehandling
- 1975: Språkbanken ("Logoteket")
- 1984: datalingvistikprogrammet
- 2000: GSLT (forskarskola i språkteknologi)
- 2004: pilotprojektet Litteraturbanken
- 2007: CLT (Centre for Language Technology)
- 2008: språkteknologi styrkeområde vid GU
- 2009: strategiska GU-medel till styrkeområdet språkteknologi
- 2011: svensk partner i META-NORD
- 2013: nationell samordnare för SWE-CLARIN

Språkbanken – vad, för vem, till vad?

vad är Språkbanken?

- ▶ en nationell resurs sedan 1975
- ▶ en FoU-enhet i språkteknologi (med nationella och internationella samarbeten, t.ex. EU-projekten META-NORD och CLARIN)
- ▶ fri tillgång till sökning i digitala, förädlade språkresurser (svenskt skriftspråk av alla genrer från alla tider):
 - ▶ textkorpora (enspråkiga och parallella)
 - ▶ lexikonresurser (moderna och historiska)
- ▶ unik kompetens inom området svenska språkresurser

(traditionellt) för vem och till vad?

- ▶ språkforskare (nordister och lingvister)
- ▶ lexikografer
- ▶ språkteknologiforskare
- ▶ utbildning
- ▶ allmänheten

[In English](#) [Anpassa](#) [Login](#)

Språk
BANKEN



GÖTEBORGS
UNIVERSITET

[Om oss](#) [Resurser](#) [Forskning](#) [Publikationer](#) [PhD program](#) [Personal](#)

SÖK

[Webbkarta](#)

Språkbanken

Språkbanken är en forskningsenhet vid Institutionen för svenska språket, Göteborgs universitet. Vårt huvudområde är språkteknologi, som handlar om hur man kan få datorer att hantera mänskligt språk i alla dess former. Vår forskning handlar om att utveckla språkteknologi för svenska språket genom tiderna. Som en del i denna forskning skapar och tillgängliggör vi språkresurser för forskare och allmänheten. På vår webbplats kan du ge dig ut på en unik och spännande språklig resa, genom att söka i våra stora svenska textsamlingar med hjälp av verktyget [Korp](#) och i svenska elektroniska lexikon med hjälp av verktyget [Karp](#).

[Läs mer...](#)

Nya ord i SALDO

2014-02-03 19:32:

[gå på ett ut göra det samma göra varken till eller från hugga i sten straffsats straffvärde svart⁴ ta illa upp ta väl upp till sista andetaget upp i daqen ur bruk visa vägen](#)

[Visa fler](#)

Exempel från Korp

böra användas, om hvar-ken sfterskildt försvårande eller särskildt förmlidande omständigheter föreligger eller om de väga Jemt, hwaremot den i andra rummet satta [straffsatsen](#), derest den är lägre än den första, kommer till användning vid öfvervägande förmlidande och, derest den är högre, vid öfvervägande försvårande omständigheter.

[runeberg-urdaqkron](#)

Forskning

[SweFN++](#)



Det svenska frasnätsprojektet

[Kulturomik](#)



Mot kunskapsbaserad storskalig kunskapsutvinning ur svensk text

[SweCxn](#)

Det svenska konstruktion-konjektet

[Diabase](#)

Språkteknologi för historiska resurser -- mot en diakronisk BLARK

[Koala](#)

Korps lingvistiska annotationer: att utveckla en infrastruktur för text-baserad forskning med högkvalitativa annotationer

[Fler forskningsprojekt...](#)

Resurser



[Korp](#)

[Användarhandledning](#)

Sök i korpusmaterialen

antal korpusar	160
token (total)	1 711 278 707
antal meningar	119 449 334



[Karp](#)

[Användarhandledning](#)

Sök i de lexikala resurserna

antal lexikon	22
ingångar (totalt)	692 727



[Lärka](#)

Lär Språket via KorpusAnalys

[Annoteringslabbet](#)

Korps annoteringslabb

[Statistik](#)

Korpusstatistik för samtliga material i Korp.

[SALDO](#)

Semantiskt och morfologiskt lexikon för språkteknologi

Ansikten utåt 1: Korp

Moderna | Parallella | Fornsvenska | Litteraturlbanken | Spf 1800-1900 | Äldre finlandssvenska | Färöiska | Digidaily

Svenska | English 



146 korpusar valda — 1 808 399 306 token

Sökhistorik

Enkel Utökad Avancerad

Sök efter Sök även som ☐ förled ☐ efterled och ☐ skiftlägesoberoende

Relaterade ord

förlora tappa_bort förlorande förspilla ge_tillspillo mista tillspillo borttappa till_spillo oförlorad släva_bort tillspilloge
ge_till_spillo

lida_nedetlag förlora förlorare förlorande ge_opp oförlorad bita_l_gräset ge_upp uppge

KWC: sortera inom korpus på: förekomst Statistik:

KWIC Statistik Ord bild

Antal träffar: 117 476

Föregående 1 2 3 4 5 6 7 8 9 10 11 .. 4699 4700 Nästa Visa kontext

ÅBO UNDERRÄTTELSE 2012

De större partierna har att bokföra den som en oundviklig **förlust**.

Efter den knappa **förlusten** mot tjeckerna kan Finland

Korpus

Åbo Underrättelser 2012

textattribut

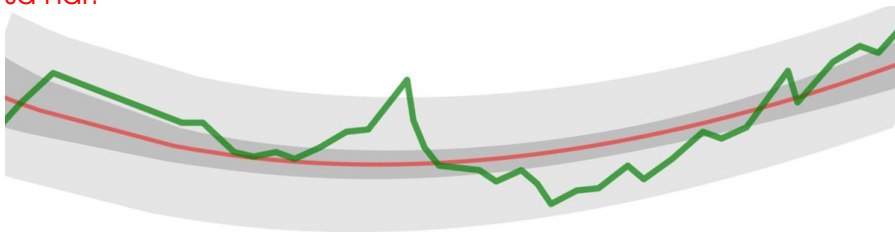
datum: 2012-09-28

ordattribut

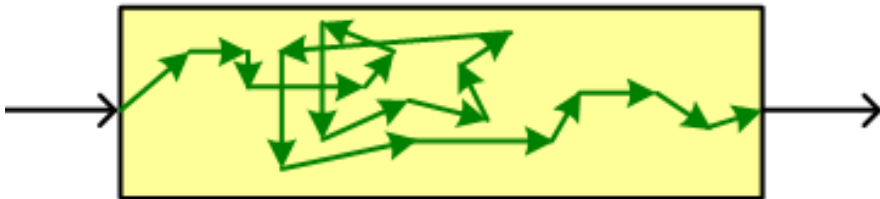
- ▶ Korp är Språkbankens korpusinfrastruktur
- ▶ Korp har en **teknisk sida** och en **användarsida**
- ▶ De tekniska lösningarna ska vara bra för användarna i stort och på lång sikt,
- ▶ vilket innebär en balansgång
 - ▶ Att bygga solida tekniska lösningar som är generella tar ibland lång tid
 - ▶ medan en särlösning för ett individuellt fall kan åstadkommas relativt snabbt

rörelsen framåt är viktig

så här:



men inte så här:



- ▶ Den viktigaste tekniska lösningen i Korp och dess syskon är
 - ▶ att korpussökmaskineriet är strikt separerat från de program som använder det, inklusive själva sökgränssnittet
- ▶ Vi talar om
 - ▶ Korps bakända och
 - ▶ dess framända
- ▶ Det betyder att man kan ha ett godtyckligt antal gränssnitt för olika grupper och olika behov och
- ▶ "användaren" är typiskt inte en människa, utan ett datorprogram



- ▶ Nästa viktiga tekniska lösning har att göra med "ingångarna" till bakändan
- ▶ Det gäller att hitta rätt frihetsgrad/abstraktionsnivå
- ▶ för då kan man blanda och ge på ett väldigt produktivt sätt
- ▶ Kanske man bäst tänker på bakändan som en samling funktioner som man kommer åt genom ett standardiserat gränssnitt.

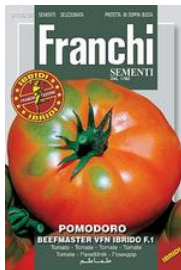
abstraktionsnivå/frihetsgrad



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT





- ▶ När man börjat tänka så
- ▶ blir det naturligt att göra så många funktioner som möjligt tillgängliga på samma sätt
- ▶ inte bara söksystemet, utan även korpusimport- och -annoteringsfunktionerna
- ▶ Man tänker förhoppningsvis mer i termer av modularisering och återanvändning
- ▶ **MEN** detta arbetssätt kräver en mycket hög och solid teknisk kompetens

- ▶ Detta finns nu i Korp:
 - ▶ KWIC-visning
 - ▶ tidsuppmärkning och funktioner för att använda den
 - ▶ annotationer: ordklass/msd, lemgram, dependenssyntax(, ordbetydelse)
 - ▶ statistikfunktioner (tabell, tårtdiagram, trenddiagram)
 - ▶ ordbild
 - ▶ bortåt två miljarder ord moderna texter, och nästan en miljard ord äldre textmaterial (i Korplabbet)
 - ▶ nedladdningsbara "meningsmängder" (slumpvis omkastade texter)
 - ▶ möjlighet att lösenordsskydda korpusar och funktioner för användaradministration
 - ▶ all mjukvara (bakända och framända) fri och nedladdningsbar för egen installation



146 korpusar valda — 1 808 399 306 token

Enkel Utökad A

Sök efter

KWIC: träffar per sida: 25 ▼

1850 1900 1950 2000

Markera alla
Avmarkera

- ☒ Akademiska texter (2)
- ☒ August Strindberg (2)
- ☒ Finlandssvenska texter (52)
- ☒ Skyddade korpusar (0)
- ☒ Medicinska texter (12)
- ☒ Skönlitteratur (5)
- ☒ Sociala medier (21)
 - ☒ Bloggmix (17)
 - ☒ Diskussionsforum (1)
 - ☒ Twitter (3)
- ☒ Tidningstexter (36)
- ☒ Tidskrifter (1)
- ☒ Dramawebben (demo)
- ☒ LäSBarT – Lättläst svenska och barnbokstext
- ☒ PAROLE
- ☒ Psalmboken (1937)
- ☒ SNP 78–79 (Riksdagens snabbprotokoll)
- ☒ SUC 2.0
- ☒ SUC 3.0
- ☒ SALT svenska-nederländska

CLT

Visa dependensträd

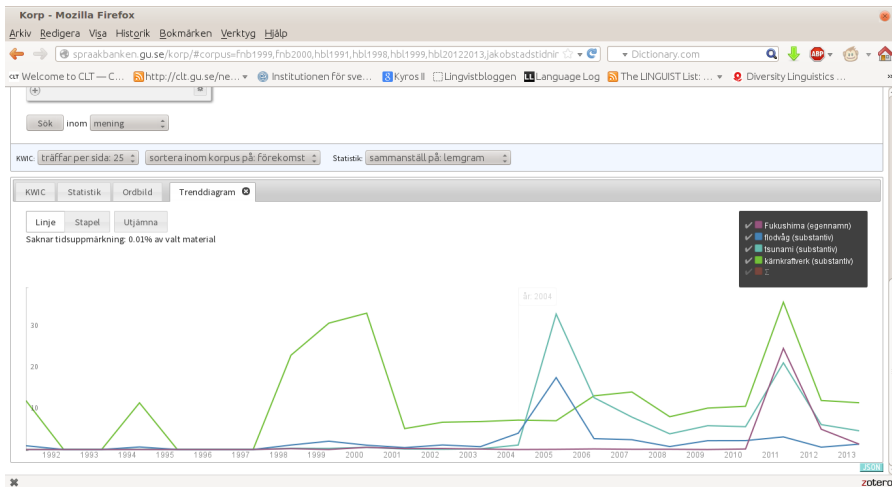
Korp: Trenddiagram



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT



Korp: Ordbild – surfa (verb)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Subjekt	surfa	Objekt	Adverbial
1. du	40 ☞	1. —	40598 ☞
2. treåring	24 ☞	2. porr	55 ☞
3. anställda	30 ☞	3. nät	32 ☞
4. svensk	33 ☞	4. våg	23 ☞
5. kund	25 ☞	5. 3g	18 ☞
6. proc ²	21 ☞	6. stund	29 ☞
7. hälft	15 ☞	7. timme	37 ☞
8. användare	15 ☞	8. nätbutiker	8 ☞
9. besökare	14 ☞	9. sida	33 ☞
10. folk	31 ☞	10. hemsida	23 ☞
11. procent	22 ☞	11. internet	20 ☞
12. emanuelkarlsten	12 ☞	12. datavolym	6 ☞
13. mp3-bok	5 ☞	13. psl	6 ☞
14. man	26 ☞	14. skånelängor	6 ☞
15. våg	9 ☞	15. 9gag	6 ☞
		1. på nät	250 ☞
		2. på internet	97 ☞
		3. bland blogg	45 ☞
		4. lite	117 ☞
		5. på sida	73 ☞
		6. på hemsida	65 ☞
		7. utomlands	47 ☞
		8. på våg	25 ☞
		9. på facebook-konto	16 ☞
		10. i shop	20 ☞
		11. på hemnet	20 ☞
		12. på Internet	26 ☞
		13. i cyberspace	20 ☞
		14. på blogg	43 ☞
		15. stund	48 ☞

















































































Korp: Ordbild – förlust (subst.)



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Preposition	Attribut	förlust	Efterställt Attribut		Förlust	verb	Verb	förlust	
1. efter	5699 	1. rak	4057 	1. på krona	5079 	1. innebära	717 	1. redovisa	2808 
2. med	7544 	2. stor	5345 	2. på dollar	1060 	2. bli	1442 	2. göra	3076 
3. trots	1363 	3. tung	1324 	3. före skatt ²	908 	3. uppgå	505 	3. lida	767 
4. utan	1641 	4. ekonomisk ²	1178 	4. före skatt	908 	4. vara	4140 	4. täcka	601 
5. mot	797 	5. ekonomisk	1178 	5. efter finansnetto	658 	5. svida	194 	5. innebära	715 
6. i och med	98 	6. snöplig	289 	6. på miljon	714 	6. komma	1041 	6. orsaka	389 
7. dagen efter	75 	7. knapp	403 	7. på euro	628 	7. betyda	221 	7. vända	391 
8. över	405 	8. hedersam	199 	8. på match	505 	8. redovisa	159 	8. vända ²	391 
9. vid	702 	9. svår	503 	9. för period	309 	9. var än	171 	9. inkassera	196 
10. på grund av	116 	10. bitter	247 	10. på miljard	323 	10. beräkna	142 	10. skylla	232 
11. även om	34 	11. enorm	272 	11. för kvartal	335 	11. bero	160 	11. tillfoga	126 
12. förutom	68 	12. klar	255 	12. av människoliv	129 	12. landa	91 	12. ta	765 
13. nära	66 	13. smärtsam	132 	13. av mångfald	147 	13. landa ²	91 	13. medföra	176 
14. till följd	24 	14. eventuell	220 	14. på mark ²	204 	14. göra	409 	14. undvika	205 
15. efter	6 	15. oväntad	155 	15. av arbetstillfälle	110 	15. kännas	158 	15. visa	348 
		16. raka	10 						
		17. raka	10 						
		18. rak	10 						
		19. sovjetisk	3 						
		20. snöplig	1 						

Ansikten utåt 2: Korplabbet

Moderna | Parallella | Fornsvenska | Litteraturlbanken | Spf 1800–1900 | Äldre finlandssvenska | Färska | Sibirientyska | Köpingsresor | Runebergtidskrifter | Bibelstallet | Lagrummet | Digidaily | Historiskt



5 korpusarvalda (alla) — 23 536 445 token

Enkel Utökad Avancerad

Sök efter Sök även som ☐ förled ☐ efterled och ☐ skiftelagesoberoende

Relaterade ord

domsrätt *extrajudiciell* spörätt rättsförfogandeinskrivning panträtt rätteligen krigsrätt förfoganderätt dispositonsrätt älsktsfrihet
förhandlingsrätt rättslig besvärsmätt rättsvis församlingsfrihet besittningsrätt självtakt rättsvis statsrättslig processrätt yttranderätt
lösningrätt åtalsrätt familjerätt nödrätt återköpsrätt äganderätt *exterritorial*rätt rättsförfogande

[Visa fler](#)

Kwic: Statistik:

Kwic Statistik

Antal träffar: 7 336

... [Visa kontext](#)

DANSSTEMNINGAR

ledd närmast af skäligen enkla taktiska hänsyn, **proklamera** de lösen: allmän **rösträtt** .
Utän allmän **rösträtt** ingen » nationell samling » säga de rösträttslösa likt själföfväldiga pojkar, som vilja mutas för att vara snälla,
Man tanke bara på alla de fäfänga ord som tala des om **rösträtt** och samling i det svenska » frasemas hus » under rösträttsdebatten anno 1906 — och man hoppas, att det ä
Hvilken är då för dem den reella betydelsen af allmän **rösträtt** — om man försöker se fullkomligt nyktert på saken?
Men hvilken glädje får då industnarbetaren af den allmänna **rösträtten** ?

Korpus

Diverse teckningar

textattribut

titel: Det Nya Sverige
datum: 1907

ordattribut

grundform:
rösträtt

Ansikten utåt 3: Korps anoteringslabb

ORP Anoteringslabbet

Ladda exempel i:

[Drama](#)

[Alla sidor](#)

[Talbanken](#)

[Läsart](#)

[Exempelkorpus](#)

Språk:

[Svenska](#)

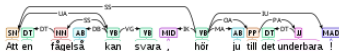
[English](#)

1. Att en fågel så kan svara, hör ju till det underbara!

Visa avancerade inställningar

Kör!

VB: Verb



ord	msd	lemma	lex	saldo	prefix	suffix	ref	dephead	deprel
Att	SN	att	att_sn.1	att.2			01	08	SS
en	DT. UTR. SIN. IND	en	en_al.1	den..1, en..2			02	03	DT
fågel	NN. UTR. SIN. IND. NOM	fågel	fågel_nn.1	fågel..1	få..av.1, få..vb.1	gel_nn.1	03	05	SS

Ansikten utåt 4: Karp

KARP

23 lexikonresurser valda (alla) — 693 410 ingångar

Sökhistorik

Enkel Utökad Editor

hund (substantiv) Sök ☒ via diavot

Träffar per sida: 25

"Förleden Wecka ar en liten swartragguer **Hund** för agaren bortkommen, eho den upragit, behagade tistalla densamma uti Madame Florins hus pa KonuMgatan."

Fullständiga träffar (41) Statistik

Föregående 1 2 Nästa

Saldo Sa... Sm... Sv... Sk... Pa... Ke... Uff... Wo... Lexin Delins ordbok Schytte... Södena... Södena... Di...

Saldo (2)

Betydelse	Lemgram	Ordklass	Primär	Sekundär	Barn (primära)	Barn (sekundära)
hund	hund (substantiv)	substantiv	djur	sällskap	<ul style="list-style-type: none">Karo –Pompe Karl XIIbandhund binda	<ul style="list-style-type: none">andeögon pannaapportera hämtabussa³ anfalla
	(236388)				Visa alla (33)	Visa alla (44)
hund ²	hund (substantiv)	substantiv	usling			
	(236388)					

Ansikten utåt 5: Lärka














Helt automatisk

Resultatsamlare

Övningsstyp	Korrekt/totalt
Lingvister/SYNT1, självstudier	3/4

Öva satsdelar

Välj en korrekt satsdel till den markerade frasen

Num	Mening	Ditt svar	Rätt svar	Länkar
4	Vattnet tas in och släpps ut i Öregrundsgrepen .	Välj relation		  
3	Barnbidraget betalas fr.o.m. kvartalet efter det då barnet föddes .	adverbial		 adverbial  
2	Färre barn skaffar man sig först då man ser en utväg ur sin fattigdom .	indirekt objekt		 indirekt objekt  

SweFN++

Sök i SweFN++

Publikationer

Utvecklingsversion

Dokumentation

Historik

Felrapport

FM-SBLEX

FrameNet Workshop 2013



NoDaLiDa 2013 workshop

Flerordsworkshop 19/3 2013

SweFN++

Lars Borin, [Dana Dannélls](#), [Markus Forsberg](#), [Karin Friberg Heppin](#), [Richard Johansson](#), [Dimitrios Kokkinakis](#),
[Leif-Jöran Olsson](#), [Maria Toporowska Gronostaj](#), [Jonatan Uppström](#), [Kaarlo Voionmaa](#).

Följ utvecklingen via  [RSS](#).

Svenskt frasnät++ (SweFN++)

Detta projekt finansieras av VR/RFI 2011-2013 (dnr 2010-6013) samt med särskilda medel från Göteborgs universitet till styrkeområdet språkteknologi (2009-2015).

SweFN++-projektet handlar om att skapa en central infrastrukturkomponent för svensk språkteknologi, nämligen en stor fritt tillgänglig lexikonresurs med rik lingvistisk information. Man kan säga att den planerade resursen kommer slå en bro mellan det förflutna och framtiden:

Det förflutna, därför att vi vill återanvända en rad fria lexikonresurser som har tagits fram i olika projekt vid olika tidpunkter av olika forskargrupper, men som sen har fått mindre användning än de förtjänar främst på grund av idiosynkratiska format och brist på driftsmedel för att underhålla resurserna;

framtiden, därför att vi till de integrerade befintliga resurserna vill lägga den typ av avancerad och mycket användbar semantisk och syntaktisk information om orden som man finner i det engelska Berkeley FrameNet (BFN) och några få liknande resurser för andra språk, ett arbete som vi planerar att göra i samarbete med den forskargrupp som står bakom BFN.

Eftersom dessa befintliga lexikonresurser representerar stora insatser i möda och pengar och eftersom de i många fall innehåller högvärdig språklig information, vill vi alltså rädda så mycket som möjligt av dem från förgängelsen samt vidareutveckla dem.

Finansierat av VR/RFI

Cure mod

domän	Med
kärnelement	Affliction Body_part Healer Medication Patient Treatment
periferielement	Degree Duration Manner Motivation Place Purpose Time
exempel	<ul style="list-style-type: none"> ▪ [Salvan]Medication [läker]LW [[skrubbsår]Affliction och [brännsår]Affliction]Affliction . ▪ [Läkaren]Healer [botade]LW både [[ryggskott]Affliction och [vatten i knät]Affliction]Affliction , innan hon hamnade i svårigheter efter att ha utfört ett föryngringsexperiment med patienten. ▪ [Genterapi]Treatment [botade]LW [dödssjuka i cancer]Patient . ▪ [Läkaren]Healer [botar]LW [kroppen]Body_part och [filosofen]Healer [botar]LW [själen]Body_part , men det krävs ett engagemang för att lyckas. ▪ Traditionellt har [stafylokockinfektioner]Affliction enkelt [botats]LW [med antibiotika]Medication . ▪ Syftet med terapi för en personlighetsstörning är inte att [fullständigt]Degree [bota]LW [patienten]Patient , eftersom det varken är möjligt eller eftersträvaransvärt. ▪ Han ansåg även att [Gud]Healer hade [helat]LW [honom]Patient [från cancer]Affliction. ▪ [Transplantation]Treatment kan ha [botat]LW [[hiv-smittad]Patient]Affliction .
lus	vb bota¹ hela¹ läka² nn läkning¹ botande¹ helande¹ av botlig¹
kommentar	Ny ram jämfört med BFN. Den ursprungliga tolkningen av ramen Cure i BFN ges här en snävare tolkning som implicerar att ett positivt resultat av någon form av medicinsk behandling föreligger.
skapad av	MTG
skapad	2012-04-02
ändrad	2013-12-09

Digital areal linguistics

Word lists

Workshop October 2010

Workshop October 2011

Project Activities

Word Lists

The languages are shown with their names and ISO 639-3 codes in parentheses. In case a language has no ISO 639-3 code, nothing is displayed.

Languages

- Hindi (ISO 639-3 code: hin)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Marathi (ISO 639-3 code: mar)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Kotsigari (no ISO 639-3 code)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Telugu (ISO 639-3 code: tel)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Bengali (ISO 639-3 code: ben)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Punjabi (ISO 639-3 code: pan)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Khasi (ISO 639-3 code: kha)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Tamil (ISO 639-3 code: tam)
LWT: [html](#) [tab-txt](#)
IDS: [html](#) [tab-txt](#)
- Nepali (ISO 639-3 code: nep)
- Gujarati (ISO 639-3 code: guj)
- Kannada (ISO 639-3 code: kan)
- Nako Kinnauri (no ISO 639-3 code)
- Sangla Kinnauri (ISO 639-3 code: kfk)
- Tibetan (ISO 639-3 code: bod)
- Kharis (ISO 639-3 code: khr)

A (revised) Swedish IDS/LWT (ISO 639-3 code: swe) list is available for download in LMF format [here](#). It can also be searched online through [Karg](#).

License



This work is licensed under a [Creative Commons Attribution 3.0 Unported License](#).

Finansierat av VR. Ett samarbete mellan Språkbanken/Göteborg, lingvistik/Uppsala och lingvistik/MPI-EVA, Leipzig

Culturomics

Culturomics: core NLP technologies

Culturomics: language over time

Culturomics: publications

Culturomics: question answering

Culturomics: text processing in historical texts

Culturomics: text processing in social media

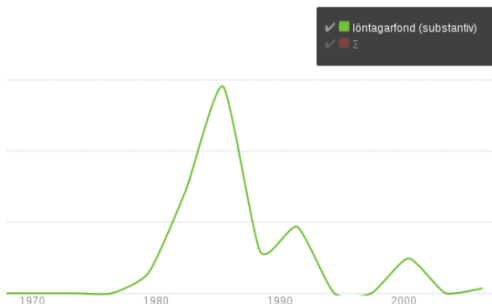
Culturomics: tracking semantic change

Culturomics: visualization

Culturomics: meetings

Exploring language over time

The following figures show the kind of results that emerge directly from a linguistically annotated text material available through Språkbanken's general corpus infrastructure. Unlike the culturomics work referred to earlier, the diagrams show the distribution of the lexemes (lexicon words) tsunami and flodvåg in a newspaper material covering the years 2001–2011, including all inflectional forms and all compounds containing these words. This is made possible by the lexical analysis tools based on handcrafted resources used for annotating Språkbanken's corpora.



Finansierat inom VR:s ramprogram Det digitaliserade samhället – igår, idag, imorgon.
Ett samarbete mellan Språkbanken/Göteborg, datavetenskap/Chalmers och datavetenskap/Lund.

kulturomik: telefoner i Sverige

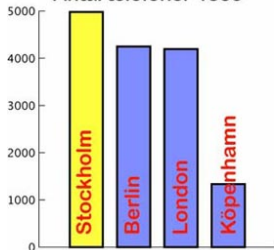


GÖTEBORGS
UNIVERSITET

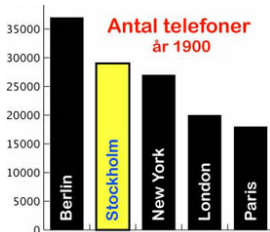
Språk
BANKEN

CLT

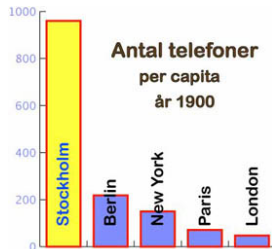
Antal telefoner 1885



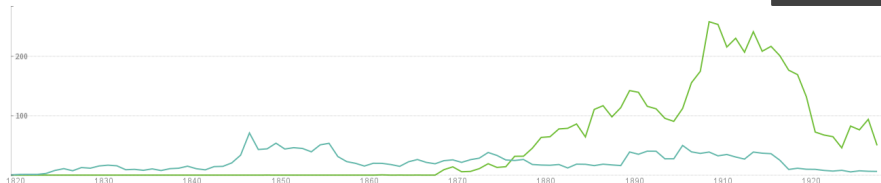
Antal telefoner
år 1900



Antal telefoner
per capita
år 1900



Linje Stapel Utjämna



Infrastruktur

Korp	>
Karp	>
Lärka	>
SBLEX	
FM-SBLEX	
Koala	

Koala – Korps lingvistiska annotationer

Projektet Koala -- Korps lingvistiska annotationer -- handlar om att utveckla en infrastruktur för text-baserad forskning med högkvalitativa annotationer.

Korpusinfrastrukturen Korp på Språkbanken (<http://spraakbanken.gu.se>) innehåller stora mängder text av olika typ och ålder, som används av forskare inom olika områden och av allmänheten. Texterna innehåller lingvistisk uppmärkning, annoteringar, som ordklasser och syntaktiska roller, vilka hjälper till att filtrera sökresultaten för användaren. De låter oss hitta "sjöng" och "sjungit" när vi söker efter "sjunga" och alla ställen där Caesar är objekt till verbet besegra utan att vi behöver titta på dem där han är subjektet, samt att vi inte behöver titta på meningar om lokaler när vi letar efter "lounge", utan kan fokusera på förekomsterna som handlar om djuret. Annoteringarnas kvalitet är avgörande för att få bra sökresultat, särskilt för forskare som annars kan behöva gå igenom tusentals irrelevanta meningar.

Målet för Koala-projektet är att förbättra annoteringarna, som har skapats automatiskt med välkända språkteknologiska metoder. Det görs genom att lägga till språklig kunskap i systemet via de många resurser som finns tillgängliga via Språkbanken, samt genom att kombinera de olika annoteringsverktygen för lexikal analys, ordklassstaggning, betydelsedisambiguering och syntaktisk analys till ett högkvalitativt system där annoteringar på ord- och meningsnivå informerar varandra och där systemet inte fattar beslut innan det har all tillgänglig information. De data och verktyg som blir resultatet kommer att göras fritt tillgängliga.

Projektet finansieras 2014–2016 av Riksbankens jubileumsfond.

Finansierat av RJ/infrastruktur

Forskning

Infrastruktur >

SweFN++ >

META-NORD >

KELLY

Kulturomik: >

CONPLISIT >

Digital areallingvistik >

ITG

MOLTO

PINCORE >

A System Architecture for ICALL >

Akademiska ordlistor

Corpus-driven induction of
linguistic knowledge

MAPIR

Svenska språket under medeltiden, fornsvenska (ca 1225-1526), finns bevarat i manuskript, brev och tidigt tryck. Dessa dokument är värdefulla för många olika forskare, såsom lingvister intresserade av svenska språkets förändring under den tiden, juridikforskare som vill undersöka medeltida lagar, teologer som studerar tidiga översättningar av bibeltexter, eller medicin-historiker som är intresserade av medeltida folkläkekonst.

I MAPIR-projektet -- Metoder för automatisk Analys av Text i digitala Historiska Resurser -- skapar vi verktyg för automatisk lingvistisk analys av fornsvenska. Projektet är relaterat till Språkbankens satsning på historiska resurser, [Diabase](#), och ligger inom forskningsområdet datalingvistik, vetenskapen om datamaskinell språkbehandling och datorstödd språkforskning. Genom att lägga till grammatisk information i digitaliserade fornsvenska texter underlättar vi studier av detta kulturarv och möjliggör nya sätt att undersöka det.

Att utveckla verktyg för fornsvenska är en utmanande forskningsuppgift, även med de främsta datalingvistiska metoderna. Detta beror på egenskaper i de fornsvenska texterna. För det första förändrades språket under den fornsvenska tiden vad gäller till exempel ordföljd och ordböjning. För det andra fanns ingen rättstavning i dagens bemärkelse. Samma ord kunde stavas på flera olika sätt. Ordet "mapir", som betyder man eller människa, stavades till exempel även "mæpr", "mander" eller "meper". Man kan till och med se olika stavningar för samma ord i ett enda stycke. För det tredje skiljer sig språket mycket åt mellan texterna. Det har gått 300 år mellan de äldsta och de yngsta texterna, och de kommer från olika geografiska områden och är av olika typ. För det fjärde kräver de flesta automatiska metoder antingen en mycket detaljerad datamaskinell beskrivning av ett språk, eller en större mängd text som redan har lingvistisk uppmärksamhet som datorn kan lära sig av. Inget av detta finns i dagsläget för fornsvenska. Kärnan i MAPIR-projektet är att utforska sätt att hantera dessa utmaningar i det fornsvenska materialet.

nya projekt: distributionella metoder



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

CLT

Forskning

Infrastruktur >

SweFN++ >

META-NORD >

KELLY

Kulturomik >

CONPLISIT >

Digital areallingsvistik >

ITG

MOLTO

PINCORE >

A System Architecture for ICALL >

Akademiska ordlistor

Corpus-driven induction of
linguistic knowledge

Corpus-driven induction of linguistic knowledge

The project aims to find automatic, corpus-based methods for inducing linguistic constructions and semantic frames, and representing their meaning using distributional semantics. In addition, the project will study the interaction between the automatically induced meaning representations and symbolic, knowledge-based resources such as frame and construction inventories, and use the representations in natural language processing (NLP) tools. It will combine two recent developments in unsupervised NLP: distributional methods for building and processing geometric meaning representations from corpora, and unsupervised semantic frame and role induction.

The results of the project will advance research in NLP and have practical benefits in applications: Corpus-induced semantic representations will be able to move beyond single words, and be formalized in terms of frame semantics and construction linguistics. Automatic syntactic and semantic analysis tools can be made more robust since they can use linguistic information beyond the word level. Linguistic resource building will benefit by the automatic methods for construction and frame discovery that the project will devise. NLP applications such as information extraction, opinion mining, grammar checking, and computer-assisted language learning can integrate the semantic frames and linguistic constructions discovered by the project, and use their distributional representations to understand their meaning.

The project is funded by the Swedish Research Council, grant 2013-4944, *Distributional Methods to Represent the Meaning of Frames and Constructions*, and lasts between 2014 and 2018.

Staff:

- [Richard Johansson](#)

Finansierat av VR

- ▶ CLARIN: ESFRI-förberedelsefas 2008-01 – 2011-06
- ▶ 9 svenska medlemmar (varav 2 partners)
- ▶ CLARIN ERIC startade 29/2 2012 med 9 medlemmar
- ▶ SWE-CLARIN-ansökan beviljad av VR 2013.
- ▶ Mål för SWE-CLARIN:
 1. bilda en svensk nod i CLARIN ERIC:
 - ▶ Göteborgs universitet (Språkbanken, SND)
 - ▶ KTH (TMH)
 - ▶ Linköpings universitet (NLP-lab)
 - ▶ Lunds universitet (Humanistlaboratoriet)
 - ▶ Stockholms universitet (datorlingvistik)
 - ▶ Uppsala universitet (datorlingvistik)
 - ▶ Språkrådet
 - ▶ DigiSam
 2. bygga en basinfrastuktur för CLARIN i Sverige



- ▶ e-vetenskap – i form av språkteknologi som forskningsverktyg – för discipliner där text (och tal) är primärdata:
 - ▶ humaniora
 - ▶ samhällsvetenskap
 - ▶ (vissa sorters) medicin
- ▶ CLARINs betydelse växer i takt med digitaliseringen av kulturarvet och den elektroniska kommunikationens utbredning

Ökat intresse för gamla gruvor

Publicerat: måndag 02 juli 2007 kl 10:22, [Nyheter P4 Norrbotten](#) |  Dela ▼



Prospektering.

Ny och effektivare teknik har gjort att intresset för gamla nedlagda gruvor har ökat markant. Lavergruvan inom Älvsbyns kommun är ett sådant exempel. Hos Bergstaten som handlägger prospekterings- och gruvfrågor ser man en stor anhopning av undersökningstillstånd i anslutning till gamla fyndigheter.

Precis som vid gruvbrytning, kräver stora mängder 'informationsglest' digitalt text- och talmaterial effektiv teknik för sökning, korrelering och korsindexering i det språkliga innehållet – även mellan språk – för att forskningen ska få ut användbara primärdata ur det.

Men bara som man kan fråga får man svar, så planerna för Språkbanken handlar om att kunna erbjuda nya sorters svar:

- ▶ korpusjämförelser
- ▶ namntagging
- ▶ textmetadata
- ▶ syntaktisk sökning
- ▶ sökvisualisering (t.ex. trender, kartor)
- ▶ smartare träffgruppering, t.ex. visningssortering efter 'semantisk' kontext
- ▶ bättre syntaxanalys
- ▶ annotering av historiska material
- ▶ talspråk och ljud
- ▶ även annan forskning än språkvetenskap
- ▶ korpvarieteter (användningar och användare), men även andra gränssnitt (med gemensamma nättjänster i bakändan)

Vi också gärna veta vilka frågor forskare och andra vill kunna ställa till materialet.

tack!

